R code for Data Science for Beginners

Day 5: Individual Exercise

Graham Jones

2025-09-20

Clean up your workspace

```
rm(list=ls(all=TRUE)) # delete all objects in the environment
cat("\014") # clear the console screen
```

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

Load the data

```r
world_data = read.csv("~/Downloads/world.csv")
```

Remember that Quarto uses a relative path so always save your data in the same folder (or under the same folder) with your Quarto code.

Democracy and female representation

Do democratic countries (`democ_regime == "Yes"`) have better female representation than non-democratic countries? Please answer this question by showing some graphs to assess the relationship between these two variables.

Initial data clean up

Analyze y

```r
summary(world_data $ women09) # look at summary statistics for female representation
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00    9.70   15.55   17.18   22.95   56.30      11
```

Analyze x

```
summary(world_data $ democ_regime) # look at summary of regime type (Yes/No)
```

```
   Length     Class      Mode
      191 character character
```

Note: There are some missing values. Labels are not intuitive. So, we will deal with these two first.

Create a smaller data set that omits NA observations.

```
women_data <- world_data[ is.na(world_data $ women09) == FALSE &
                          is.na(world_data $ democ_regime) == FALSE, ] # keep only rows with
```

Let's make sure that we did this correctly.

```
summary(women_data $ women09) # check again, no missing values
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    9.85   15.55   17.23   22.73   56.30
```

```
summary(women_data $ democ_regime)  # check regime variable again
```

```
   Length     Class      Mode
      178 character character
```

We can see that NA cases have been correctly removed.

Re-labeling the values

```
women_data $ regime_label <- factor(women_data $ democ_regime,
                                   levels = c("Yes", "No"),
                                   labels = c("Democracy", "Autocracy"))

summary(women_data $ regime_label) # check the new labels
```

```
Democracy Autocracy
      111        67
```

```
rm(world_data) # The original dataset is no longer needed.
```

Q1: Describe Y (numerically and graphically)

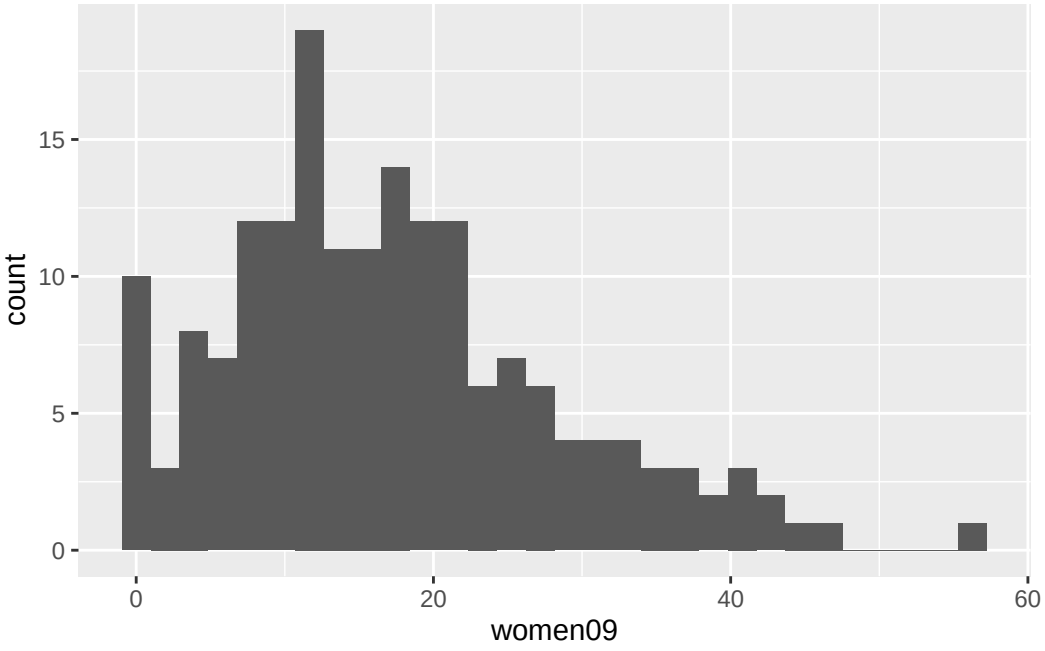
Numerical summary

```
summary(women_data $ women09) # numerical summary of female representation
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    9.85   15.55   17.23   22.73   56.30
```

Graphical summary

```
ggplot(women_data, aes(x=women09)) +
  geom_histogram(bins = 30) # histogram of female representation
```



Q2: Describe X (numerically and graphically)

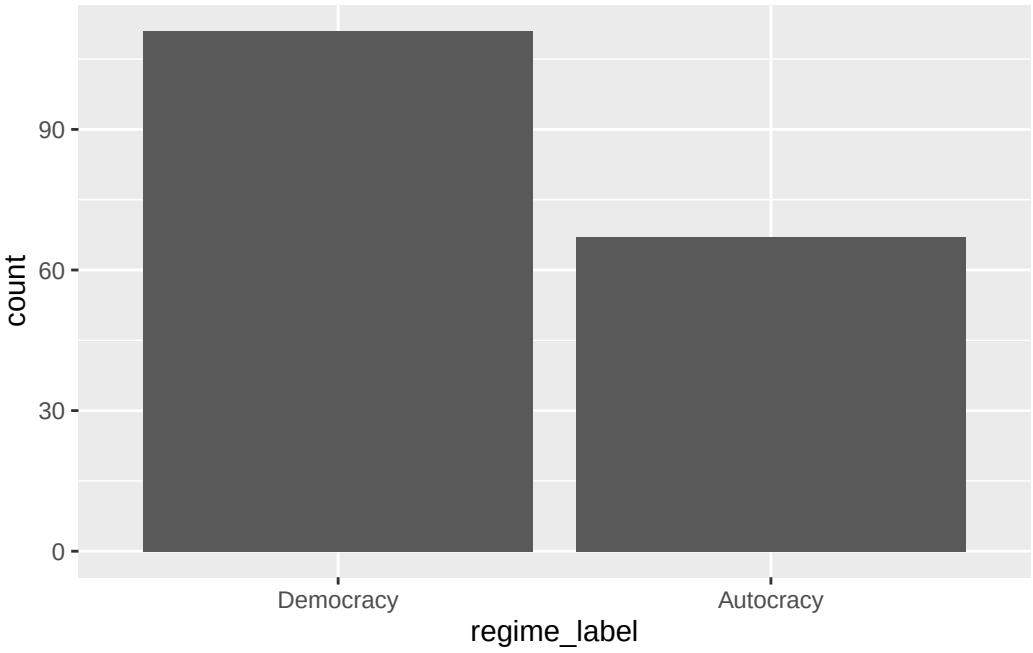
Numerical summary (frequency table)

```
summary(women_data$regime_label) # frequency count of regime type
```

```
Democracy Autocracy
      111        67
```

Graphical summary (bar chart)

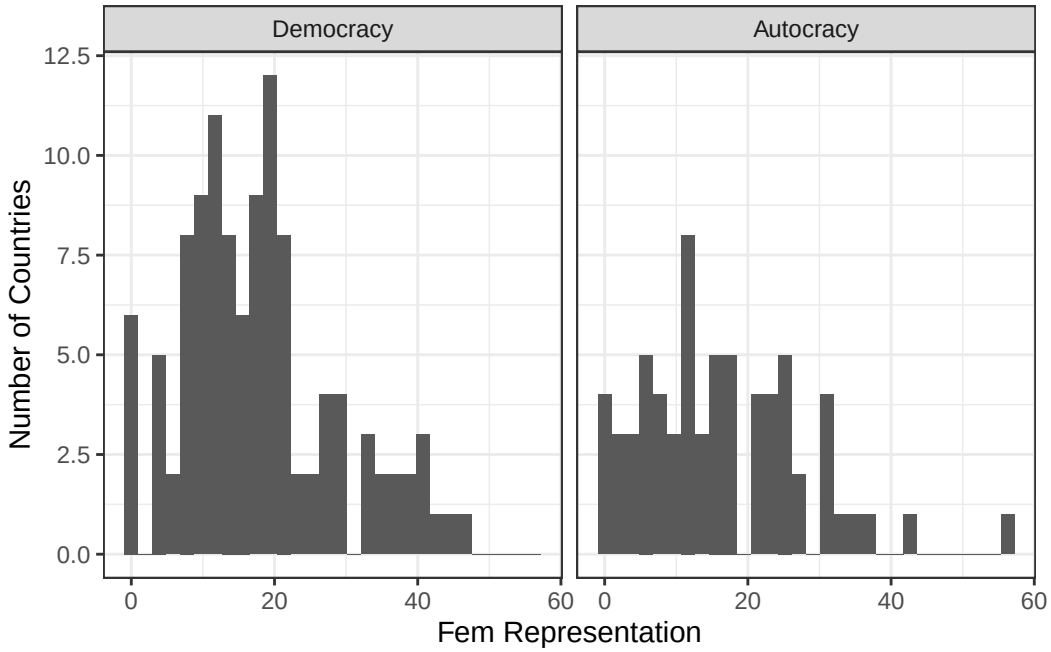
```
ggplot(women_data, aes(x = regime_label)) +
  geom_bar()  # bar chart of countries by regime type
```



Q3: Describe X-Y graphically

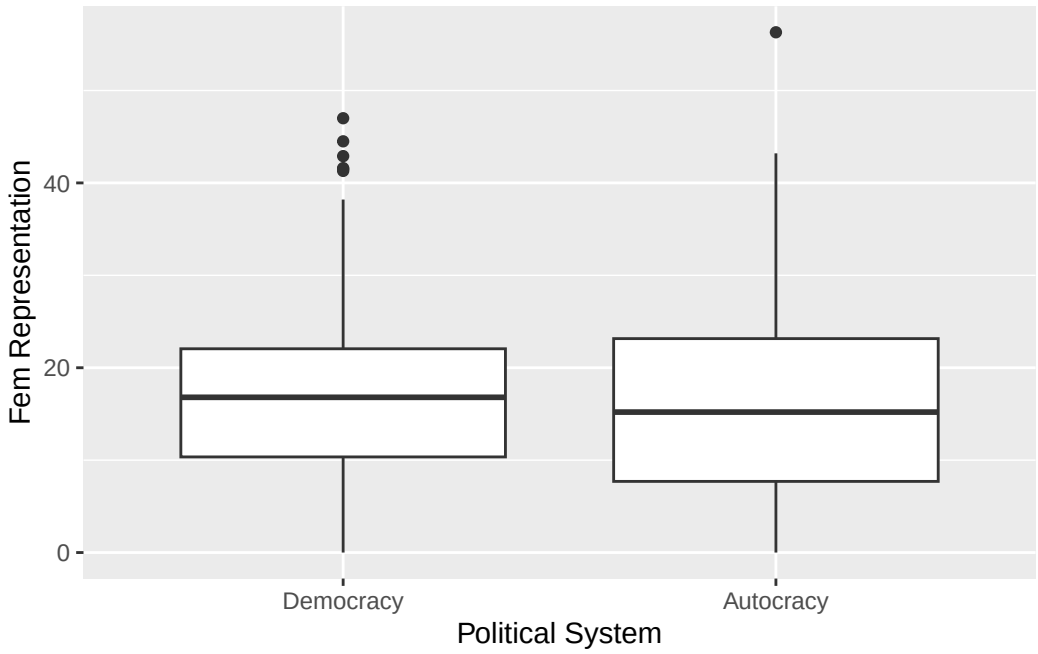
Histograms

```
ggplot(women_data, aes(women09))+
  geom_histogram(bins=30) +
  theme(axis.text.x = element_text(size = 14)) +
  xlab("Fem Representation") +
  ylab("Number of Countries") +
  theme_bw() +
  facet_grid(.~regime_label) # histograms split by regime type
```



Box-plots

```
ggplot(women_data, aes(x = regime_label, y = women09))+
  geom_boxplot() +
  xlab("Political System") +
  ylab("Fem Representation") # boxplot comparing democracies and autocracies
```


Fem Representation
40
20
0
Democracy
Autocracy
Political System